



Towards Insider Summarization for Mediation Instead of Moderation: Examining Wikipedian Views on Key Elements of Discussion Summaries

SOOBIN CHO, Human Centered Design & Engineering, University of Washington, USA

MARK ZACHRY, Human Centered Design & Engineering, University of Washington, USA

DAVID W. MCDONALD, Human Centered Design & Engineering, University of Washington, USA

In Wikipedia, conflict plays a vital role in refining knowledge. Yet some disputes persist without clear resolution or rule violations, resulting in long, circular discussions without moderator intervention. These deter neutral and third-party editors and often end only when one side gives up. In such cases, what is needed is not moderation but mediation—support that helps participants voice perspectives and reach consensus. However, mediation is hard to implement manually, as it requires summarizing complex discussions with a full grasp of both content and the Wikipedia community's culture¹. One potential solution is delegating this task to AI—but what content should be included in summaries that mediate and facilitate disagreements? Through a three-phase interview study with 14 Wikipedians, we examined how they read and interpret discussions, create their own summaries, and evaluate large language model (LLM)-generated summaries presented as technology probes. Our findings show that Wikipedians expect summaries to include key discussion elements—usernames, topics, arguments, sources, editor behavior, rule violations, and resolution—with particular emphasis on community-related context that reflects insider understanding. Based on these insights, we apply a new theoretical framework grounded in computer-mediated communication (CMC) theories such as common ground theory, warranting theory, and social presence theory to examine Wikipedians' interpretations of user identity. We also discuss design implications for discussion summaries on Wikipedia talk pages.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; *Empirical studies in HCI*.

Additional Key Words and Phrases: Wikipedia talk pages, Dispute mediation, Discussion summary elements, Technology probe

ACM Reference Format:

Soobin Cho, Mark Zachry, and David W. McDonald. 2025. Towards Insider Summarization for Mediation Instead of Moderation: Examining Wikipedian Views on Key Elements of Discussion Summaries. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW500 (November 2025), 25 pages. <https://doi.org/10.1145/3757681>

1 Introduction

In Wikipedia, a space where people collaboratively build knowledge, disagreement and conflict are inevitable—and not inherently negative. Disagreements can result in valuable processes that help refine article content. While many disputes can be resolved through established protocols,

¹When we refer to Wikipedia's culture, we mean the community norms and practices, many of which are influenced by its policy environment, including the Five Pillars [44], policies, guidelines, and essays. Throughout this study, 'culture' will encompass all of these elements.

Authors' Contact Information: Soobin Cho, soobin30@uw.edu, Human Centered Design & Engineering, University of Washington, Seattle, USA; Mark Zachry, zachry@uw.edu, Human Centered Design & Engineering, University of Washington, Seattle, USA; David W. McDonald, dwmc@uw.edu, Human Centered Design & Engineering, University of Washington, Seattle, USA.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM 2573-0142/2025/11-ARTCSCW500

<https://doi.org/10.1145/3757681>

others persist without significantly violating any rules. These often lead to prolonged discussions between disagreeing parties, because in Wikipedia “winning [...] is simply about staying in the discussion long enough” [20].

As these discussions continue, the barrier to entry increases for those not closely following, who must catch up on an ever-lengthening exchange—often discouraging neutral or third-party editors [13, 20]. Yet it is precisely these voices that help move conflicts forward; without them, discussions tend to go in circles and only end when the less committed party gives up and leaves. This often leads to contributors leaving Wikipedia entirely [20], as their efforts to participate are overlooked. If such conflicts were resolved through simple moderation—similar to other online communities—the outcome would likely be the same, as their willingness to contribute would still go unrecognized despite their active participation.

What is needed in these situations is mediation through the support of a facilitator—someone who “aids consensus reaching [...] by encouraging the participants to voice” their perspectives, by appropriately mirroring the community’s discussion [20]. However, introducing an ad hoc facilitator is extremely challenging. Unlike moderation, which is triggered by clear violations of protocol, mediation requires catching up on the entire conflict, understanding both the content and context, and offering summaries that reflect a deep, insider grasp of how the community engages in discussion. Having human facilitators consistently perform this role across numerous disputes in real time is simply not feasible. But this could be a role that AI might be able to play.

Our goal is to investigate what content should be included in summaries that can mediate and facilitate disagreements. Specifically, we examine what types of information from discussions should be included to help Wikipedia editors effectively grasp situations on article talk pages. To achieve this, we explore how Wikipedians 1) interpret and make sense of discussions, 2) generate their own ideal summaries after this interpretive process, and 3) evaluate a large language model (LLM)-generated summary that is minimally customized to the community—used as a technology probe—against their expectations.

In summary, our study asks the following research questions:

- (1) How do Wikipedians read and comprehend discussions on Wikipedia?
- (2) How would Wikipedians summarize discussions on Wikipedia?
- (3) What do Wikipedians expect from discussion summaries on Wikipedia?

Our study involved interviews with 14 Wikipedians, conducted in three phases. In the first phase, participants selected one of four real discussions from Wikipedia talk pages. They were asked to read and interpret the discussion as if they were viewing it directly on Wikipedia, using a think-aloud method to verbalize their thoughts. In the second phase, participants wrote a summary of the discussion, assuming they were creating it for other Wikipedians. In the third phase, participants were shown a minimally community-customized LLM-generated summary, used as a technology probe to elicit their expectations for a discussion summary. They were then asked to evaluate the summary and reflect on the potential utility of a discussion summarization tool.

Through the interviews, we uncovered what Wikipedians need and expect from discussion summaries. They wanted key elements from the discussion to be included, and emphasized the importance of community-related elements for deeper understanding. These key elements included usernames, discussion topics, editors’ arguments, cited sources, editors’ behaviors, rule violations, and discussion resolution. Their significance emerged in both the interview’s first phase, when participants read the discussions, and the second phase, where participants generated their own summaries. The importance of community-related elements—usernames, cited sources and policies, editor behavior, and discussion resolution—was further reinforced in the third phase, where participants evaluated the LLM-generated summaries.

Our study offers four main contributions. First, based on our findings, we discuss how Wikipedians' interpretations of editors' user identities—through usernames and user pages—relate to theories of computer-mediated communication (CMC). Second, we propose design implications for creating discussion summaries on article talk pages, including which elements should or should not be included, and the potential need for multiple summaries to reflect different user needs and contexts. Third, our three-phase interview method triangulates key elements of discussion summaries from three distinct perspectives, helping to uncover and validate what is important. Finally, this study is the first to thoroughly investigate how Wikipedians read, process, and interpret talk page discussions.

The following sections begin with background and related work that situate our study within broader research on discussions in Wikipedia and online communities. We then describe our three-phase interview method, followed by findings from each phase. Finally, we discuss theoretical implications of Wikipedians' understanding of editors' user identities. We then explore design implications for discussion summaries and reflect on our methodological approach.

2 Background and Related Work

In this section, we present the background and related work, organized into two areas: discussions in Wikipedia and discussions in online communities.

2.1 Discussions in Wikipedia

In examining discussions in Wikipedia, we review related work that 1) emphasizes the constructive role of disagreement, 2) outlines common discussion types, 3) identifies key characteristics of their dynamics, and 4) examines how readers engage with discussions.

2.1.1 Discussions—or Disagreements—are Valuable. Discussions happen when people disagree, and this is no different on Wikipedia. As a platform for collaboratively building knowledge, discussion is not only expected—it's essential. Jemielniak, both a researcher and active community member, noted that Wikipedia is more oriented toward disagreement than collaboration itself [20, 21]. In fact, most editors interact with others only when they disagree or need help [20].

Previous research on Wikipedia disputes suggests that disagreements—even when they escalate into heated conflicts—can be constructive [23, 24]. Productive rebuttals, such as counter arguments and refutations, are linked to improved outcomes [24]. In this sense, discussions, or disagreements, and even disputes are not inherently negative; they are meaningful processes that help refine Wikipedia content.

2.1.2 Types of Discussions in Wikipedia. Wikipedia hosts a variety of discussions, differing in content, participants, and style. Examples include the Administrators' noticeboard, where administrators handle exclusive tasks like user conduct and closing formal discussions [42]; Requests for Adminship (RfA), which invites the community to comment on candidates for administrator status [5, 10, 47]; Requests for Comment (RfC), a formal process to resolve disputes by inviting community feedback [19]; and Articles for Deletion (AfD), where editors debate the deletion, merging, or movement of articles [19, 35]. In addition to these dedicated discussion pages, a wide range of administrative and technical pages have talk pages that cover metadata (e.g., categories, templates), sub-groups or teams (e.g., WikiProjects) and community guidelines (i.e., policy).

Article Talk Page Discussions

Every encyclopedic article on Wikipedia also has a dedicated talk page for content discussions [25, 39]. Here, we define a “discussion” as a single thread initiated by an editor's section header and description, along with all subsequent replies. These discussions are displayed in chronological

order without categorization, so overlapping topics may emerge if editors are unaware of earlier threads.

Early in Wikipedia, the standard editing practice was to propose changes on the talk page before making edits [39]. After some discussion and a reasonable waiting period, the change could be made, and the talk page would be updated to reflect this. This older, “talk before you type” approach has changed with time. A more current approach is Bold, Revert, Discuss (BRD): an editor makes a bold edit—an encouraged practice where editors directly fix problems instead of just talking about them—and if another editor disagrees, they revert the edit. The reverted or reverting editor then initiates a discussion on the talk page to reach consensus [49].

Article talk page discussions vary in topic and intensity, with some leading to difficulty in reaching consensus or escalating into broader disputes. Depending on the nature of the dispute, resolution paths may include seeking a Third Opinion (WP:3O), posting to relevant noticeboards [43], initiating a Request for Comment (WP:RfC), using the Dispute Resolution Noticeboard (WP:DRN), requesting Administrator Attention (WP:RAA), or, in more serious cases, escalating the matter to the arbitration committee (WP:ArbCom) [51].

Given this complexity and potential contentiousness, our study focuses on article talk page discussions.

2.1.3 Characteristics of Wikipedia Discussions. Prior studies emphasize that understanding Wikipedia discussions requires familiarity with the community’s culture. They also elaborate on the typical lifecycle and consequences of conflicts not governed by formal protocols, as described in the introduction.

Community Culture in Wikipedia Discussions

Many studies have highlighted how Wikipedia discussions are embedded in community culture, often emphasizing the importance of Wikipedia policies as a framework created by Wikipedians to facilitate collaboration through shared language and standardized strategies [3, 25, 38]. These policies and procedures are also associated with reduced conflict [22]. For instance, studies on AfD discussions show that policy-based arguments are common, and validated arguments demonstrate knowledge of policies and community values [34, 35]. Studies on article talk pages similarly found that referencing Wikipedia guidelines is frequent [32, 33, 39], enough to reflect community concerns and work patterns over the long term [3]. Even when arguments are not directly related to policy, they gain strength when appealing to community values [27]. These studies underscore the importance of familiarity with Wikipedia culture for understanding discussions on the platform. Our study supports these prior findings and illustrates the value of understanding the context of Wikipedia and specific user activities before designing and developing a technology to support Wikiwork.

Characteristics of Conflict in Wikipedia Discussions

Wikipedia has well-developed procedures to moderate conflict. Over the past 24 years, it has established numerous protocols and introduced roles like administrators to help manage disputes. Conflicts occur daily on the platform, and many are resolved through these rules [20]. For example, behaviors that contribute to conflict—such as negative messaging [12] or excessive bickering [19]—are moderated when they violate community guidelines like ‘No Personal Attacks’ [46] and ‘Harassment’ [45].

However, many conflicts persist beyond the scope of protocol-based moderation, and these unresolved cases often stretch on indefinitely. One reason is that editors involved in such disputes are resistant to changing their opinions [53]. As Jemielniak notes, “when the rules and sources do not give one side the advantage, winning an argument is simply about staying in the discussion long enough” [20]. As discussions grow longer, it becomes increasingly difficult for others—especially

neutral or third-party editors—to join or rejoin the conversation, since doing so requires reading a lengthy chain of back-and-forth arguments [13, 20]. These drawn-out discussions often end only when the less persistent party gives up. Those who surrender not only lose the argument but often leave Wikipedia entirely, which may ultimately harm article quality [20].

Importantly, the very participants who are discouraged by the length of these conflicts—neutral and third-party editors—are often the ones who can help move the dispute forward [20]. A qualitative study on edit wars² found that such conflicts are typically resolved through external intervention [53]. Here, simply moderating the conflict to halt the disagreement can mirror the outcome of a surrender—where one side’s contributions are dismissed entirely. What’s needed is not just moderation, but mediation and facilitation that support constructive disagreement.

Our study examines how Wikipedia editors read and interpret discussions to determine what should be included in summaries. By understanding their reading practices, we ultimately aim to support mediation—not just moderation—through summaries that offer a clear Wikipedian view of ongoing disagreements.

2.1.4 Reading and Comprehending Discussions in Wikipedia. Few studies have examined Wikipedia reader behaviors, and even fewer have focused on readers of discussions. In 2014, Okoli et al. conducted a systematic literature review of 99 studies on Wikipedia readership [29]. They categorized these studies by factors such as knowledge domain, publication type, and page type. Their analysis showed that only two studies specifically focused on article talk pages, and three when including other discussion types. Other studies have indirectly explored the role of talk pages for Wikipedia readers. For instance, Elmimouni et al. developed a trust taxonomy for assessing Wikipedia articles, noting that readers’ evaluations may involve examining contributor discussions [11]. Additionally, Antin et al. demonstrated that reading, particularly talk pages, is a key activity through which newcomers learn about Wikipedia and its editing practices [2]. However, no studies have explored the cognitive processes involved in how Wikipedians read discussions. Our study aims to investigate how Wikipedians start to comprehend difficult discussions to understand which information elements are important to comprehension.

2.2 Discussions in Online Communities

In this section, we 1) introduce some computer-mediated communication (CMC) theories relevant to online discussions, 2) review studies on reader behaviors in discussion contexts, and 3) present work on tools designed to support discussions in online communities.

2.2.1 CMC Theories and Online Discussions. Various computer-mediated communication (CMC) theories have been applied to online discussions. Among them, common ground theory, warranting theory, and social presence theory share a focus on viewing communication as socially situated and on the interpretive practices through which users make sense of one another’s messages. These theories offer a useful lens for explaining how Wikipedians interpret user identities from usernames and user pages when reading article talk page discussions—we later elaborate on this point in the Discussion section.

According to Clark and Brennan, common ground refers to the shared knowledge and assumptions that interlocutors rely on and update through grounding—providing evidence that messages are understood well enough for current purposes [7, 8]. This theory has been linked to Wikipedia disputes [4], argumentation-based collaboration [15], and tools designed to support online debate [18].

²Edit warring refers to the situation where editors repeatedly remove each other’s contributions, often due to strong disagreements [50].

Warranting theory, as developed by Walther and Parks in the context of computer-mediated communication, posits that the more information is perceived to be immune from manipulation by the person it describes, the more influential it becomes in shaping others' impressions [41]. Walther later applied this theory to participatory websites, such as social networking sites and wikis. These sites include incidental aggregate user representations (AURs)—such as friend counts or views—derived from behaviors not intended as signals. Because such data are beyond the control of any single user, they may carry strong warranting value [40].

Social presence theory, first defined by Short et al., describes the perceived presence and relational salience of others in mediated communication [14, 36]. While often studied in online learning, Cortese et al. explored its role in expressing opinions within synchronous, text-based environments [9].

2.2.2 Reading and Comprehending Discussions in Online Communities. Like studies on Wikipedia, most research on discussions in other online communities focuses on analyzing past discussions rather than on how community members read them. Few have examined the real-time behaviors of readers—what information they notice, how they process it, and the cognitive steps involved.

Pian et al. used eye-tracking to study how users read online health discussions, showing that information need, urgency, and text length affect skimming and deeper reading behaviors across summary and detailed views [30]. Although not a community platform, Zhang et al., through interviews with group chat users, found that people often skim with scrolls, skip less important content, and search for specific information when catching up on past conversations [54].

Our study focuses on how readers, as members of their own online community, process discussions—what information they catch and how they make sense of it in context.

2.2.3 Tools for Discussions in Online Communities. Helping users quickly and easily make sense of a discussion is the first step toward mediating it. Many studies have designed tools specifically to support this sense-making process in online discussions, often by using summarization techniques [1, 28, 55]. Other approaches include organizing comments into topics [16], displaying opinions along a spectrum from pro to con [26], and visualizing relationships between discussants [37].

In online community contexts, most of these tools aim to support information extraction or conflict moderation. For example, Ren et al. developed a system that summarizes Q&A threads on Stack Overflow to assist readers in identifying useful answers [31]. Choi et al. designed a moderation tool for Discord that visualizes conversational metrics—such as activity and toxicity levels—to detect and respond to emerging issues early [6].

Our study, while not yet designing a tool, takes a different direction. We use summarization not as a means of extracting information or moderating conflict, but to explore how discussions might be better mediated and facilitated. Unlike moderation or extraction, mediation requires a deeper, insider understanding of how community members engage with discussions. To inform this, we examine how readers within a community interpret ongoing conversations—what they notice, how they process it, and what enables them to participate meaningfully.

3 Method

To understand what kinds of informational elements Wikipedians want in discussion summaries, we conducted a semi-structured, three-phase interview with 14 Wikipedians, with the final phase involving a technology probe—an LLM-generated summary of a selected dispute from a Wikipedia article talk page.

This study was reviewed and approved by the Institutional Review Board (IRB) at our university. We also adhered to the Wikimedia Foundation's (WMF) research guidelines for conducting research on WMF-related websites. A research project page was created on Meta-Wiki to describe the

project's details and facilitate community review. The page outlined the research background, questions, scope, participant recruitment and incentives, methodology, and potential study impacts. During the research, the lead author responded to questions from community members.

3.1 Creating Technology Probe Summaries

To use in the third phase of our semi-structured interviews, we developed an LLM prompt and generated summaries of selected discussions from Wikipedia talk pages to serve as a technology probe. Technology probes are prospective designs or prototypes intended to prompt participants to consider how a technology might align with their context, goals, needs, or desires. They are often designed to fill conceptual gaps participants may have about how a technology could work or to help problematize aspects that are difficult to articulate without direct interaction [17].

Our goal was not to create a prompt and output that fully solved the summarization problem, but rather to offer a probe that would encourage participants to reflect on their expectations for a summary and to envision the potential value of a discussion summarization tool. We ensured the summaries were reasonably reliable—avoiding hallucinations—through iterative prompting, but intentionally limited additional customization to the Wikipedia community to avoid preemptively aligning the summary with community-specific expectations.

We selected four discussions, one each from the following Wikipedia articles: Erotica, February 29, Barber Paradox, and French Language. In choosing these, we avoided highly contentious or sensitive topics that might trigger strong personal biases and interfere with understanding participants' general cognitive processes during reading. At the same time, we ensured that each discussion involved conflict with potential for escalation, presented multiple perspectives, and met a minimum level of complexity. Each was also short enough to be carefully read and reflected on within the limited time of the interview.

Our prompt evolved through five iterations, resulting in: "*This is the content in question. [content]* *This is a discussion containing a dispute about what to write in a Wikipedia article. [discussion]* *Identify the dispute. Be very concise.*" The process of identifying the dispute, refining the prompt, and the summaries of the four discussions are detailed in [Appendix A](#).

3.2 Recruitment

For recruitment, we compiled a list of candidate study participants from Wikipedia's Dispute Resolution Noticeboard (DRN) volunteers, administrators, and active editors. Candidate editor accounts were screened for editing activity by reviewing their public edit history. We included editors with over 200 article edits and 50 talk page edits since 2022, ensuring they had been active over the past two years and could offer perspectives rooted in insider knowledge.

Editors who met our criteria were contacted using Wikipedia's "Email this user" feature. Our invitation email included the research objective, participant identification method, research procedures, a link to the Meta-Wiki project page, and a participation questionnaire. The questionnaire collected participants' gender, age, editing experience, frequency of encountering disputes, and interest in participating in an interview. It also provided the article and discussion titles of four sampled disputes and asked participants to select at least one to discuss during the interview.

Over a two-month period, we contacted 21 DRN volunteers, 14 administrators, and 110 active editors, totaling 145 Wikipedians. Of these, 21 completed the questionnaire, and 14 participated in interviews. Recruitment ceased upon reaching data saturation, where recurring insights were noted.

Interview participants varied in roles and ages, with most having over five years of editing experience. Notably, while most contacted users were active editors, recruitment resulted in an even ratio of administrators to editors. Table 1 presents descriptive information and the article

Table 1. Descriptive Information and Discussion Topics of the Interview Participants

| ID | Age | Role | Editing Experience | Discussion Topic |
|-----|---------|--------|--------------------|------------------|
| P01 | 18 ~ 24 | editor | 5 or more years | Barber Paradox |
| P02 | 18 ~ 24 | editor | less than a year | French Language |
| P03 | 25 ~ 34 | admin | 3 to 5 years | French Language |
| P04 | 25 ~ 34 | editor | 5 or more years | February 29 |
| P05 | 18 ~ 24 | editor | 3 to 5 years | February 29 |
| P06 | 55 ~ 64 | admin | 5 or more years | Barber Paradox |
| P07 | 65 ~ 74 | admin | 5 or more years | February 29 |
| P08 | 35 ~ 44 | admin | 5 or more years | French Language |
| P09 | 35 ~ 44 | admin | 5 or more years | French Language |
| P10 | 25 ~ 34 | admin | 5 or more years | Barber Paradox |
| P11 | 25 ~ 34 | editor | 5 or more years | Barber Paradox |
| P12 | 45 ~ 54 | editor | 5 or more years | French Language |
| P13 | 25 ~ 34 | editor | 5 or more years | Barber Paradox |
| P14 | 55 ~ 64 | admin | 5 or more years | Barber Paradox |

subjects chosen by interview participants. Although gender information is excluded in the table to protect participant anonymity, 10 identified as male, 2 as female, and 2 chose not to disclose.

3.3 Interview Procedure

Interviews were conducted via video conferencing. At the beginning of each interview, participants made a final selection of one discussion to review from four subjects: Erotica, February 29, Barber Paradox, and French Language. The article and discussion titles were presented as they appeared in the questionnaire to assist with selection.³ Once a discussion was chosen, participants received the discussion's PDF file via chat, along with another PDF illustrating the "Difference Between Revisions," which highlighted relevant article sections and edits.

During the first phase of the interview, participants engaged in think-aloud sessions, verbalizing their thoughts while reading the selected discussion to help reveal their cognitive processes. They were asked to review the discussion as if they had encountered it on Wikipedia, and then share their thoughts, reading approach, and past experiences with similar discussions. In the second phase, participants wrote their own summaries of the discussion and shared them via chat. They were instructed to write for other Wikipedians, with no restrictions on content or length. Once they had shared their summary, participants were asked to explain the important aspects of their summary. In the last phase, participants were presented with the LLM-generated summary of the same discussion via screen sharing and asked to read and share their thoughts on it. Participants also discussed how they envisioned a summary generation tool being used by Wikipedians, its potential applications, and its limitations.

Interviews lasted 70 to 120 minutes, with most around 90 minutes. At the end of each interview, participants could choose one of three like-minded organizations for a \$25 donation as a token of appreciation. All interviews were audio recorded with participants' consent and transcribed for analysis.

³We note that while the majority of participants chose all four subjects in the questionnaire, there was one specific subject that none of them were actually willing to discuss during the interview: the dispute from the talk page of the Erotica article.

3.4 Analysis

We collected interview transcripts and participant-generated summaries of their chosen discussions. Interview transcripts were analyzed using thematic analysis and memo generation to develop a comprehensive understanding. The first author iteratively identified themes and concepts, merging memos into groups based on relevance and similarity. Following each iteration, all authors discussed and refined the themes to accurately represent the data. This process yielded distinct concepts across three areas: how Wikipedians read discussions, the role of talk pages, and the role of summaries.

Participant-generated and LLM-generated summaries were analyzed using open coding for thematic analysis. The first author conducted iterative coding on seven participant summaries to develop a preliminary codebook, which was refined through multiple discussions among all authors. The first author then coded the remaining seven summaries, updating the codebook as needed. Once all participant summaries were coded, further discussions were held to ensure accuracy. Finally, the three LLM-generated summaries were coded using the refined codebook. This approach allows us to compare participant-generated summaries with LLM-generated ones using the same analytical dimensions, uncovering important similarities and differences.

4 Findings

Through the interviews, we uncovered what Wikipedians need and expect from discussion summaries. In the first phase, participants identified key discussion elements: usernames, discussion topics, editors' arguments, cited sources, editors' behaviors, rule violations, and discussion resolution. The second phase reaffirmed the importance of these elements through their reappearance in participant-generated summaries. The third phase emphasized the value of community-related elements—username, cited source and policy, editor behavior, and discussion resolution—based on participants' evaluations of the LLM-generated summaries.

4.1 Discussion Elements Revealed Through the Reading Process

The think-aloud sessions revealed participants' reading process, highlighting seven key information elements that they glean from discussions. Participants made sense of and assessed these elements within the community's culture, though the conclusions they drew varied greatly, even among experienced users. In the following sections, we outline how participants processed each element.

4.1.1 Username. Talk page discussions consist of comments with the contributor's editor ID attached at the end of each comment, allowing readers to identify the username of each contributor. All 14 participants checked the usernames as soon as they started reading the discussions and often explicitly mentioned them during their think-aloud, such as: "*starting with the first note by [Orange]...*" (P07) or "*now [Peach] responds...*" (P11). As such, participants consistently noted who made each argument.

At this moment, information inferred from usernames could influence participants' assessments of the arguments. As active community members, participants sometimes recognized editors by their usernames possibly from their ongoing work in Wikipedia. For example, P12 said, "*I think I recognize the name [Carrot]. I think it's [...] administrator.*" Participants also noted when comments were made anonymously, indicated by an IP address instead of a username, meaning the editor was not registered or logged in. P01 noted, "*The first comment is from an IP, which is interesting.*" Some participants mentioned that "*A lot of anonymous IP addresses are not as experienced.*" (P06)

4.1.2 Discussion Topic. All talk page discussions are centered on the modification of the article content to which the talk page is attached. The 'discussion topic' refers to the specific proposed changes, such as removing, altering, or adding content.

Participants quickly identified the discussion topic based on the title and early content. While most participants had no thoughts on the discussion topic itself, judgments were made when participants felt that the discussion missed the point. P11, while reading a discussion on Barber Paradox that included a debate about whether an editor's addition of "solution" was correct or not, said, "*They're having an argument based on content [...] Wikipedia is not about being right [...] Wikipedia is about citing reliable sources and synthesizing sources in an appropriate way.*" P12 also criticized the discussion on French Language for missing the point, saying, "*I would criticize [Blueberry] and other users for not looking at what an official language is [...] they should go in to find what the Wikipedia article says about official language and [discuss] does it meet that criteria.*"

4.1.3 Arguments and Argument Flow. Discussions on modifying content naturally include arguments about the suggested changes, such as support, opposition, and new suggestions. As participants read the discussions, they integrated individual arguments to understand the claims and responses. They continuously reevaluated the merits of each argument and its flow, ultimately taking sides.

At a surface level, participants expressed simple impressions of a comment's importance. For instance, P09 consistently judged, "*Now that's an important claim [...] That's just a repetition [...] that's not very useful [...] That's not the point.*"

Slightly deeper, judgments were made on the argument's completeness, such as whether the argument is supported by a reliable source. Experienced participants identified comments that made claims without sources, saying, "*this is an interesting addition [...] however, no citations at all.*" (P05) Participants also determined whether a statement is right or wrong, saying, "*that's probably correct*" (P09) or "*that guy is not correct.*" (P10)

Lastly, participants quickly and casually judged whether they agreed or disagreed with each comment, but their opinions differed. For instance, while reading the discussion on February 29, P07 said, "*[[Orange] says,] 'Therefore there is no choice but to break templates' [...] breaking templates is, you know, not an appropriate option [...] Okay, so this person, [Lemon], is offering an option [...] This is good.*" In contrast, P05 had the opposite reaction to the same argument: "*'Therefore there's no choice but to break templates' [...] Yeah, they're correct! [...] I would side with [Orange] [...] and disagree with [Lemon].*"

4.1.4 Cited Source. Wikipedia policy requires that statements in articles be supported by 'reliable sources.' As a result, editors on talk pages often cite sources to support their arguments. Of the three discussions, only the one on the French language included external sources, where an editor provided three sources: one from Wikipedia, one from *Vatican News*, and one from a French diplomatic website. Participants assessed the reliability of each source; for instance, P12 said, "*I would take Vatican News as being the most trustworthy source [...] Although I would say the French government is a reliable type of source, I'd be worried about taking their reliability on the extent of the French language in particular. And I would never choose Wikipedia as a reliable source.*"

4.1.5 Editor Behavior. Wikipedia emphasizes the importance of respectful and productive discussions on talk pages, with established behavioral guidelines. Accordingly, participants evaluated whether the editors in the discussions were engaging "well" based on their comments. The evaluations ranged from simply acknowledging the editors' communication style to examining whether they provided evidence for their claims and their attitude when making those claims.

Noticing the editors' communication style was evident when some participants described a comment as "*straightforward*" (P01) or "*blunt and succinct.*" (P07) The evaluation of whether the editors provided evidence aimed to ensure more mature discussions, with credit given regardless of agreement with the claims: "*This user cites policy and the other cites an essay [...] They do*

have some grounds for saying something, that is nice." (P01) Lastly, assessing the editors' attitudes usually happened when participants encountered an editor being uncivil towards another. The degree to which an attitude was considered uncivil, however, was different from participant to participant, resulting in different evaluations for all three discussions. For instance, commenting on the discussion about February 29, P04 assessed that it was "*resolved peacefully*" and all contributors "*are being civil*," whereas P05 described one editor as "*mean*" and P07 suggested another editor should "*walk away, come back [...] with a cooler head.*"

4.1.6 Rule Violation. Wikipedia enforces policies and guidelines to prevent actions that disrupt constructive discussions. In two sample discussions, rule-breaking behavior occurred: edit warring in the February 29 discussion and sock puppetry in the French Language discussion. The February 29 discussion began when an outsider summoned two editors who were edit warring, while in the French Language discussion, one editor was accused of using a fake account, known as a 'sock puppet.'

Most participants focused on these violations as they contravened community guidelines, but the extent to which they were considered serious varied among participants. For instance, participants reading the French Language discussion showed varying degrees to which they linked the violation with the claims made. Among the five participants, two participants (P02, P03) closely linked the editor's activity with how they judged their argument, saying, "*The [Kumquat] guy was a sock puppet the whole time [...] That kind of reduces his credibility by a bit because I've seen trolls do this before.*" (P02) On the other hand, both P09 and P12 did not consider sock puppetry a big deal as long as bad intention or execution was not spotted in the discussion. P09 explained, "*[If] they use the accounts and they do good things, then it's not a problem.*" One participant (P08) expressed mixed feelings, saying, "*That doesn't mean the merits of what they were saying isn't valid [...] [But] it's hard for me to detach that aspect of it because a lot of people use sock puppets to manipulate discussions.*"

4.1.7 Discussion Resolution. Since article talk page discussions focus on potential changes to an article's content, they typically lead to some form of resolution—whether it is to approve, reject, or modify the proposed changes. Resolution can be reached through consensus, or if the debate concludes without reaching a consensus, the status quo may be maintained. After reading the discussions, participants noted how they were resolved. Observing how the discussions were resolved served to understand if the discussion would require additional, possibly future, actions.

Participants also evaluated whether they agreed with the resolution of the discussion, and their opinions varied. For instance, regarding the Barber Paradox discussion, P14 "*agreed with the end result*," while P11 said, "*The right solution is somewhat in between.*" In the French Language discussion, P09 "*agree[d] with the conclusions,*" while P12 assessed that it was "*not brilliantly handled.*"

4.1.8 Beyond Discussion Text: Metadata. Participants also utilized information beyond the discussion text, including metadata on the talk page, such as format and timestamps, as well as external information like cited links and Wikipedia user pages. This metadata helped them understand the situation more comprehensively.

Participants noticed unusual formatting, such as a struck-out comment from a sock puppet, and quickly assessed the situation: "*I see a big few struck comments. I read down at the bottom [...] there was a sock puppet, meaning they've been previously blocked.*" (P03)

Timestamps were referenced to track the temporal flow and status of the conversation. The intervals between comments helped determine whether the conversation was happening in real-time: "*The timestamps are quite close together. So this is two people da da da...*" (P14) Knowing when the last comment was posted also helped participants assess whether the discussion was ongoing: "*It was two months ago, and so it's something that seems to have quite settled down.*" (P12)

Table 2. Inclusion of summary elements in participant-generated summaries and technology probe summaries. Grayed-out sections indicate that the element is not applicable to that particular discussion, as it was not present.

| Topic | ID | User name | Discussion Topic | Argument | Supp. Details | Judg. on Argument | Cited Source | Eval. of Source | Editor Behavior | Rule Violation | Violator | Executor/Moderator | Resolution | Eval. of Resolution |
|-------|-------|-----------|------------------|----------|---------------|-------------------|--------------|-----------------|-----------------|----------------|----------|--------------------|------------|---------------------|
| BP | P01 | | | o | | | | | | | | | o | |
| | P06 | | | | o | | | | o | | | | o | |
| | P10 | | | o | | | | | | | | | | |
| | P11 | | | o | | o | | | o | | | | | |
| | P13 | | | | | | | o | | | | o | | |
| | P14 | | | o | | | | | | | | | | o |
| FL | P02 | o | o | o | | | o | o | o | o | o | | | |
| | P03 | o | o | o | o | | o | o | | o | o | | o | |
| | P08 | | o | o | o | | o | o | | o | o | | o | |
| | P09 | o | o | o | o | | o | o | | o | o | o | | |
| | P12 | | | o | | | | | | | | | | |
| | P04 | | | | o | | | | | o | | | o | |
| F29 | P05 | o | o | o | | | | | | o | o | | | |
| | P07 | | | o | | | | | o | o | | o | o | |
| | BP | Probe | | o | o | o | | | | | | | | |
| FL | Probe | o | o | o | | | o | o | | | | | | |
| F29 | Probe | o | o | o | o | | | | | | | | | |

Out of 14 participants, 10 stated that they would click on every provided link. This included all five participants of the French language discussion, who indicated that they would click on the sources to verify if they were official and to check if “*what this user is saying is actually true*” (P03).

Nine participants explicitly mentioned that they would refer to the Wikipedia user page. User pages were seen as integral to understanding discussions, even though they were unrelated to the arguments themselves. Through user pages, participants gleaned insights into the editor’s experience, intentions, identity, and expertise, aiming to understand the editor as a whole. One common reason cited for checking an editor’s experience was to “*give them the benefit of the doubt*” (P08) and “*politely explain*” (P05) if they were new to Wikipedia.

4.2 Participant-Generated Summaries: Reaffirming the Elements

The discussion elements described in the previous section reappeared as summary elements in participant-generated summaries (see Table 2). However, variations arose due to differing participant opinions on which information to omit for neutrality and which details were significant enough to include.

4.2.1 Username. The inclusion of usernames varied among participants. However, in the summary of the Barber Paradox discussion, where one of the editors was an unregistered “IP” editor, half of the participants still distinguished between “IP user” and “registered user,” even while maintaining anonymity. This reflects Wikipedians’ cognitive approach of distinguishing between unregistered and registered editors, even when they do not find it necessary to mention usernames explicitly.

4.2.2 Discussion Topic. While the discussion topic was not included in every summary, it was more evident in the summaries of the French Language discussion. This is likely because the topic, “whether French is an official language of Vatican City,” is straightforward and can be answered with a simple yes or no, unlike other discussions. Notably, P12 added a point not explicitly mentioned

in the text to their discussion topic, explaining that they did so because they felt the core issues “*weren’t being really properly addressed.*”

4.2.3 Argument. Most summaries included the main arguments of the editors. However, participants differed in their opinions on which arguments were important enough to mention—some included only the initiator’s argument, some included the two main opposing views, and others included every claim. While supporting details were often considered too extensive to include, the French Language discussion included details related to the reliability of sources, leading some participants to incorporate them in their summaries. Most summaries excluded judgments to maintain neutrality despite numerous judgments that were made. However, one participant considered an action taken against an editor to be inappropriate and included this in their summary: *[IP editor added good faith contributions [...] they didn’t need to be reverted.]* (P11)

4.2.4 Cited Source. Sources were only mentioned in the French language discussion, and four out of five summaries of the discussion included cited sources. Again, participants had varying opinions on which sources were significant enough to include—one mentioned only Wikipedia, another mentioned *Vatican News* and the French diplomatic website, and the remaining two mentioned all three sources. All four summaries also contained evaluations of the sources. Most participants borrowed claims made by the editors to maintain neutrality, but one provided their own evaluation: *[He cited a Wikipedia article, which (although in another language) is still not counted as a credible source. The [Kumquat] user had two other non-Wikipedia sources, but he still lost a bit of credibility just by citing Wikipedia.]* (P02)

4.2.5 Editor Behavior. The inclusion of editor behavior varied among summaries, with some participants explicitly stating that it was “*not important.*” (P09) Mentions of behavior matched what emerged during the reading process, covering aspects such as communication style (*[... conclude the conversation in terse, short sentences...]* (P07)), the attitude of providing evidence (*[All parties stated Wikipedia policies to back their ‘argument’...]* (P13)), and negative or potentially uncivil actions (*[The breaking-editor is frustrated...]* (P07), *[...even goes as far as to revert another user’s edit, indicating a bit of defensiveness...]* (P02)).

4.2.6 Rule Violation. Most summaries of the two discussions involving rule violations mentioned that a violation had occurred. Notably, the username of the sock puppet violator in the French Language discussion appeared in all four summaries that mentioned the rule violation. Some summaries also included the username of the moderator or the person who executed the punishment.

4.2.7 Discussion Resolution. Half of the summaries included a brief note on how the discussion was resolved, such as *[A compromise was reached]* (P01) or *[no clear consensus was established]* (P04). While some participants explained that they avoided including their judgment on the resolution to maintain neutrality, one summary included a personal judgment (*[I also concur with the removal]* (P14)), with the participant explaining that they were accustomed to providing opinions as an administrator.

4.3 Technology Probe Summaries: Evaluating and Envisioning

Technology probe summaries, which focused on the discussion text, elicited participants’ thoughts and emphasized elements closely tied to community culture—username, cited source and policy, editor behavior, and discussion resolution. Participants also identified three ways in which summaries could assist Wikipedians: 1) quick content identification, 2) promotion of correct understanding, and 3) facilitation of participation.

4.3.1 Evaluation of Key Elements in Technology Probe Summaries. Participants' evaluations of the technology probe summaries could be divided into four main comments: 1) mentioning usernames has advantages and disadvantages, 2) details regarding sources and policies are lost, 3) behavioral nuances are overlooked, and 4) resolution details are missing.

Mentioning usernames has advantages and disadvantages. Participants expressed contrasting opinions on mentioning usernames. Some argued that it is important to mention names because who makes which arguments matters in Wikipedia. “*One editor argues that...’, oh, [...] Who you are does matter in Wikipedia. So knowing the identities of the people involved, that’s a big thing.*”(P11) On the other hand, some participants argued against mentioning names to mitigate bias. “*If you name drop, sometimes people will focus on the name rather than the points [...] If they recognize the name [...] it creates an unintentional bias.*”(P08) For the involved editors, mentioning names could make it too personal and increase sensitivity to how they are depicted. “*people who are involved in arguments might get very upset at even small differences [...] people [might also] get upset that they’re not included.*”(P12) Thus, mentioning usernames required complex consideration of various stakeholders.

Details regarding sources and policies are lost. Since Wikipedia is guided by specific policies and necessitates sources for claims, the absence of details on sources and policies was considered a significant omission of the community’s culture. “*This [summary] looks at it [...] without regard to policies quoted, without regard to reliable sources. That’s where I think this summary breaks down a bit.*”(P06) P13 also explained, “*Policies should have been included, as it[Wikipedia] is guided by policies.*” While the LLM-generated summary of the French Language discussion mentioned cited sources, some participants felt that details were missing. P09 noted the absence of evaluation, remarking, “*If I were to point out that the French Wikipedia was cited, I would at the same time point out that that is not a reliable source.*” P14 wanted details on whether a source was provided for a claim, even in discussions where their topic did not involve sources. “*One of the things that it[summary] is not picking up is that the person who added that sentence hadn’t added a source.*”(P14) Missing these details meant that the summary overlooked important aspects of community culture.

Behavioral nuances are overlooked. Among different aspects of behavior, participants noted the loss of emotional nuance. P07 highlighted the absence of emotional nuance and its potential impact on understanding the situation: “*This is very cut and dry [...] It certainly leaves out [...] the tone [...] It’s devoid of any kind of emotion [...] [This] could lead to misunderstanding. I think nuances in language are important.*” (P07) This concern—that the absence of emotional nuance could lead to misunderstandings—is particularly relevant, as discussions inherently involve the exchange of differing opinions.

Resolution details are missing. Lastly, participants who sought to understand the current status through a summary noted the absence of details on how the discussion was resolved. “*I think it could include some more information like the status, like what decision was ultimately reached. It doesn’t mention that.*” (P04) P11 specifically mentioned wanting to know actionable points, saying, “*If I were reading a summary, I would want [to know] if there’s an unresolved part of the conversation.*” This indicated that participants assumed they might need to join the discussion if necessary, underscoring the need for explanations about resolution, similar to those in participant-written summaries.

4.3.2 Envisioning the Potential of AI-Generated Discussion Summaries. After reading the LLM-generated summaries, participants envisioned the functional potential of an AI-generated discussion

summary. Eleven out of fourteen participants explicitly agreed on its potential value, envisioning three ways in which summaries could assist Wikipedians: 1) quick content identification, 2) promotion of correct understanding, and 3) facilitation of participation.

Summaries aid quick content identification. First and foremost, summaries were considered an efficient way of understanding a discussion. Through such efficiency, P04 expected that summaries would enhance productivity. *"I think it might allow discussions to be much more productive. Editors don't have to read through dozens of replies to get an idea of where the discussion is headed [...] They could get up to speed very quickly."*(P04) Similarly, P02 described summaries as a news report for editors. *"I'd rather, you know, read a little summary than have to read literal pages worth of arguing back and forth. So in a way the AI is kind of like doing a little news report."*(P02) As such, participants emphasized the efficiency of reading summaries over reading the actual discussion.

Summaries promote correct understanding. Summaries were also expected to encourage correct understanding by providing an unbiased and macroscopic depiction of the discussion. P05 was aware of their own potential human bias when reading a discussion and explained that a summary could be a solution. *"[If] I read the entire discussion fully, I might be missing some context or in my head overemphasizing an argument. But if a summary was there to pick out the highlights, that would help me guide my own thinking in my head of what more to pay attention to."* Though not self-reflective, P12 also pointed out that summaries can mitigate human flaws when reading a discussion. *"It might also be a way of having users [to] not pay undue attention to certain users [...] and also not give undue attention to what's at the beginning and the end of the conversation."*(P12) P12 also suggested that summaries may prevent a reader from fixating on a wrong topic that is being dealt in the discussion by providing a macroscopic view. *"A summary might be a very useful way of saying [...] that's not the job of the discussion [...] the main thing it[summary] would help [is], it would set people's views [right]."*(P12) This expectation particularly demonstrates the role of a summary in providing accurate and unbiased content.

Summaries can facilitate participation. Lastly, summaries could assist Wikipedians in deciding whether to engage in a discussion and prepare them for participation. P11 imagined that they *"could use the AI summary to see [if] this is even worth reading."* P08 also anticipated that summaries *"would be used as a jumping-off point [...] They would read this and go [...] I don't really want to get involved [...] or [...] this is in my wheelhouse, I'm interested."* If they decide to join the discussion, a summary could work as a preparation. *"If I saw this, I could go to this discussion and kind of have a good background of what's going on."*(P06) This aligns with participants' willingness to engage when necessary, as demonstrated by their comments on how the LLM-generated summaries missed details about how the discussion was resolved. It also suggests that participants expect the summaries to facilitate their transition from third-party readers to active discussion participants.

5 Discussion

In this section, we examine Wikipedia's user identity through computer-mediated communication (CMC) theories and discuss the design implications for creating discussion summaries for article talk pages on Wikipedia. We also reflect on our study method, which involved a three-phase interview using LLM-generated summaries as a technology probe, and outline our study's limitations and possible directions for future research.

5.1 User Identity in Wikipedia Discussions and CMC Theories

In Wikipedia, a community with a strong shared culture, usernames and user pages played a role beyond simply distinguishing who said what in discussions. Their deeper roles and effects can be

interpreted through computer-mediated communication (CMC) theories discussed previously in this paper.

5.1.1 Common Ground Theory. Usernames and user pages served as indicators to verify whether an editor possessed the shared knowledge emphasized by common ground theory. As highlighted throughout this paper, understanding Wikipedia's community culture is as crucial as knowledge about the discussion topic itself. Therefore, participants actively sought cues indicating whether editors involved in a discussion possessed such community-specific knowledge. This was especially evident in participants' differentiated reactions to usernames versus anonymous IP addresses. Not only did every participant note the presence of an "IP user" during their reading, but even in summaries where anonymity was maintained, some still distinguished between "IP user" and "registered user." An anonymous IP address implicitly signals that the individual might not truly belong to the Wikipedia community, suggesting a potential lack of experience and shared common ground. Participants, therefore, were concerned not merely with anonymity, but specifically with the potentially weaker common ground of anonymous editors.

5.1.2 Warranting Theory. Participants frequently referred to editors' edit counts—the total number of edits made by a specific editor—and account creation dates, both of which can be found on user pages, as concrete indicators for assessing how much common ground, or Wikipedia experience editors possessed. For example, P13 explicitly noted, "*I go to see [...] how many edits that that person has made and how long the person has been on Wikipedia.*" Such information acts as incidental aggregate user representations (AUR), which are beyond an individual's direct control and thus carry significant warranting value. In fact, beyond the scope of this interview, edit count is frequently used as a shorthand for assessing an editor's activity level within Wikipedia, to the extent that various tools exist specifically for calculating and comparing edit counts [52].

5.1.3 Social Presence Theory. User pages were also described as a means of sensing relational salience, as defined in social presence theory. Several participants mentioned that they would visit editors' user pages to gather additional information about the editors, even when that information was unrelated to the discussion topic itself. Their intent was to achieve a more holistic understanding of the editors rather than to form bias, as seen from P09's explanation. "*Sometimes it's useful to know what people say about themselves. Like which languages do they know? [...] I don't want to use them to form a biased opinion [...] but at least maybe understand their perspective. This can be useful to understand their knowledge, their background.*"(P09)

However, higher social presence may not necessarily be beneficial in the context of discussions. Usernames inherently create greater social presence compared to anonymous contributions. As several participants expressed concern, revealing usernames could unintentionally heighten social presence, shifting the focus toward who made an argument rather than the argument itself.

5.2 Design Implications for a Discussion Summary

Through our analysis, we identified elements that could be included in a LLM-generated summary, as well as decision contexts in which these elements should be included. We also extend the design implications by noting that participants suggested a need for summaries from different perspectives.

5.2.1 What Should LLM-Generated Summaries Look Like? The full list of elements identified throughout the process includes: username, discussion topic, editors' arguments, supporting details, judgment on arguments, cited sources, evaluation of sources, editor behavior, rule violations, violators, executors/moderators, discussion resolution, evaluation of resolution, and metadata such as format, timestamps, cited links, and user pages. These elements can be categorized as

mandatory elements, recommended elements, optional elements, elements to avoid, and elements with user-dependent inclusion preferences.

- **Mandatory Element:** Editors' arguments
- **Recommended Elements:** Cited sources, Evaluation of sources, Rule violations, Violators, Discussion resolution, Cited links
- **Optional Elements:** Discussion topic, Supporting details, Executors/moderators, Format, Timestamps
- **Elements to Avoid:** Judgment on arguments, Editor behavior, Evaluation of resolution
- **User-dependent Element:** Username, User pages

Mandatory elements are the minimum essential content needed to understand a talk page discussion. Recommended elements are important information noted by the majority of participants; while they might not be included in the simplest summaries, they are crucial for understanding the discussion within the community's culture. Optional elements are either too obvious (e.g., discussion topic) or too detailed and specific to include in a summary. Elements to avoid are those that can introduce bias and compromise neutrality, and thus should be omitted. Finally, as highlighted in the findings, the inclusion of usernames, as well as user page information, may vary based on user preference.

If summaries containing these elements are designed to be interactive, the interactions can support users' understanding and improve the perceived reliability of the content through cross-referencing with the original discussion. Users seeking deeper context or wanting to verify specific information could click on summary sentences to view the corresponding part of the discussion. Additionally, since Wikipedia editors frequently use hyperlinks when citing sources or referring to Wikipedia policies, it may be helpful to collect and present all mentioned links alongside the summary.

Examining these elements raises questions like, "Is there really a single right summary?" Differing contexts of potential use, as identified by our participants, indicate the need for different summaries in those differing contexts.

5.2.2 Need for Summaries from (at least) Two Different Perspectives. The main distinction between a summary containing only mandatory elements and one that includes both mandatory and recommended elements is the inclusion of information related to community culture. Depending on whether this information is incorporated, different summaries can be produced.

This difference can also be understood as the contrast between LLM-generated summaries that closely follow the discussion text and participant-generated summaries that interpret and explain the discussion within the context of community culture. When participants reviewed the LLM-generated summaries alongside their own, they perceived a disparity between the two, framing it as a distinction between insider and outsider summaries. "*I feel like the AI has kind of the knowledge of a general person trying to contribute to Wikipedia [...] as opposed to the knowledge of, you know...*" (P11) Many participants valued their own insider summaries more: "*I think mine works better for established sort of editors [...] Understanding [the discussion] fully sort of in and of itself requires a knowledge of Wikipedia and how it works.*" (P01) However, some participants responded positively to the outsider-perspective summaries: "*This is a great summary [...] It cuts out a lot of the commentary pieces [...] It's just straight to the facts, straight to what is going on [...] It allows them to see the center of the argument.*" (P10)

This divergence of opinions suggests that an "insider" summary may not always be preferable to an "outsider" summary. The type of summary desired can vary according to user needs. For instance, individuals primarily interested in the discussion topic may find summaries with too

much insider information overly detailed. A potential solution to address this could be offering both, as suggested by P05: “*I guess there can be two summaries generated.*”

5.3 Reflection on Method

For this study, we conducted a three-phase interview involving: 1) a think-aloud session while reading a discussion, 2) writing their own summary, and 3) evaluating a summary generated by our technology probe. This method effectively identified and validated the key elements for a discussion summary, triangulating our findings by confirming them from three distinct perspectives. In the first phase, we observed what information participants noticed and inferred while reading the discussion, providing a broad overview of the types of information that emerged in their cognitive processes. In the second phase, participants wrote their own summaries, allowing us to understand which information they deemed necessary and valuable to include. Comparing the first and second phases confirmed that the information participants identified while reading the discussion was considered important enough to include in their summaries. In the third phase, we presented participants with an LLM-generated summary, enabling them to envision using such a summary generation tool and to reconsider what should be included in a discussion summary from the perspective of using AI-driven assistance. Comparing the second and third phases revealed that participants’ evaluations of the LLM-generated summary were primarily concerned with the presence or absence of elements they had included in their own summaries, confirming the importance of elements appeared in the second phase. This third phase also reconfirmed that the evaluated summary elements overlapped with the information that emerged during the first phase, showing alignment between participants’ understanding of discussions and their expectations of the summary generation tool. Thus, this method validated which summary elements are crucial when generating discussion summaries from three perspectives, resulting in robust findings through a triangulating approach.

5.4 Limitations and Future Work

While our work aimed to achieve an in-depth understanding of how Wikipedians read discussions, there were limitations in fully comprehending this due to the study’s nature. Firstly, since the study relied on voluntary participation, those particularly interested in AI technology’s application to Wikipedia may have been more inclined to participate. Additionally, although participants read actual discussions, their experiences differed from real situations where Wikipedians naturally encounter and read disputes. Consequently, their reading and comprehension may vary based on the purpose and context of reading. Lastly, while we aimed to encompass various topics and cases in our sample discussions, we could not cover all possible scenarios. Therefore, different reading behaviors that might arise in cases not included in our sample were not captured. Despite these limitations, this study provides a foundational understanding of how Wikipedians read and process controversial discussions, which can enable future research. Based on this study, we plan to further explore the potential of AI tools that can assist Wikipedians in quickly understanding discussions. This could involve studying how to design prompts that provide summaries reflecting the understanding of Wikipedians more accurately or understanding how Wikipedians intend to utilize such summary generation tools as part of their community activities.

6 Conclusion

Our study aimed to determine what information Wikipedia article talk page summaries should include to be well-aligned with the community’s culture and expectations. To achieve this, we conducted three-phase interviews with 14 Wikipedians, involving: 1) a think-aloud process while reading a discussion, 2) writing their own summaries, and 3) evaluating an LLM-generated summary.

In the first phase, participants identified seven types of discussion elements: usernames, discussion topics, editors' arguments, cited sources, editors' behaviors, rule violations, and discussion resolution. The second phase reaffirmed the importance of these elements as they reappeared in the participants' summaries. In the third phase, LLM-generated summaries that are minimally customized to the community elicited participants to emphasize elements more closely related to community culture—usernames, cited sources and policies, editor behavior, and discussion resolution. We also found that participants expect summaries to help users 1) quickly identify the content, 2) promote understanding, and 3) facilitate participation. Based on our findings, we discussed theoretical implications around Wikipedians' interpretations of user identity through usernames and user pages, and design implications for creating discussion summaries for Wikipedia article talk pages. Additionally, we examined how our three-phase interview method validated our findings through triangulation.

References

- [1] Lucas Anastasiou and Anna De Liddo. 2021. Making Sense of Online Discussions: Can Automated Reports help?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI EA '21*). Association for Computing Machinery, New York, NY, USA, Article 471, 7 pages. [doi:10.1145/3411763.3451815](https://doi.org/10.1145/3411763.3451815)
- [2] Judd Antin and Coye Cheshire. 2010. Readers are not free-riders: reading as a form of participation on wikipedia. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) (*CSCW '10*). Association for Computing Machinery, New York, NY, USA, 127–130. [doi:10.1145/1718918.1718942](https://doi.org/10.1145/1718918.1718942)
- [3] Ivan Beschastnikh, Travis Kriplean, and David McDonald. 2008. Wikipedian Self-Governance in Action: Motivating the Policy Lens. *Proceedings of the International AAAI Conference on Web and Social Media* 2 (09 2008), 27–35. [doi:10.1609/icwsm.v2i1.18611](https://doi.org/10.1609/icwsm.v2i1.18611)
- [4] Matt Billings and Leon A. Watts. 2010. Understanding dispute resolution online: using text to reflect personal and substantive issues in conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (*CHI '10*). Association for Computing Machinery, New York, NY, USA, 1447–1456. [doi:10.1145/1753326.1753542](https://doi.org/10.1145/1753326.1753542)
- [5] Moira Burke and Robert Kraut. 2008. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) (*CSCW '08*). Association for Computing Machinery, New York, NY, USA, 27–36. [doi:10.1145/1460563.1460571](https://doi.org/10.1145/1460563.1460571)
- [6] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 262 (Oct. 2023), 30 pages. [doi:10.1145/3610053](https://doi.org/10.1145/3610053)
- [7] Herbert H. Clark and Susan Brennan. 1991. *Perspectives on Socially Shared Cognition*. American Psychological Association, Washington, DC. [doi:10.1037/10096-006](https://doi.org/10.1037/10096-006)
- [8] Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to Discourse. *Cognitive Science* 13, 2 (1989), 259–294. arXiv:https://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1302_7 [doi:10.1207/s15516709cog1302_7](https://doi.org/10.1207/s15516709cog1302_7)
- [9] Juliann Cortese and Mihye Seo. 2012. The Role of Social Presence in Opinion Expression During F2F and CMC Discussions. *Communication Research Reports* 29 (01 2012), 44–53. [doi:10.1080/08824096.2011.639913](https://doi.org/10.1080/08824096.2011.639913)
- [10] Katie Derthick, Patrick Tsao, Travis Kriplean, Alan Borning, Mark Zachry, and David W. McDonald. 2011. Collaborative Sensemaking during Admin Permission Granting in Wikipedia. *Proceedings of the 14th International Conference on Human-Computer Interaction (HCII'11)* (2011).
- [11] Houda Elmimouni, Andrea Forte, and Jonathan Morgan. 2022. Why people trust Wikipedia articles: Credibility assessment strategies used by readers. In *Proceedings of the 18th International Symposium on Open Collaboration* (Madrid, Spain) (*OpenSym '22*). Association for Computing Machinery, New York, NY, USA, Article 9, 10 pages. [doi:10.1145/3555051.3555052](https://doi.org/10.1145/3555051.3555052)
- [12] Dominique Fréard, Alexandre Denis, Françoise Détienne, Michael Baker, Matthieu Quignard, and Flore Barcellini. 2010. The role of argumentation in online epistemic communities: the anatomy of a conflict in Wikipedia. In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics* (Delft, Netherlands) (*ECCE '10*). Association for Computing Machinery, New York, NY, USA, 91–98. [doi:10.1145/1962300.1962320](https://doi.org/10.1145/1962300.1962320)
- [13] Vicenç Gómez, Hilbert J. Kappen, and Andreas Kaltenbrunner. 2011. Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia* (Eindhoven, The Netherlands) (*HT '11*). Association for Computing Machinery, New York, NY, USA, 181–190. [doi:10.1145/1995966.1995992](https://doi.org/10.1145/1995966.1995992)

- [14] Charlotte N. Gunawardena. 1995. Social presence theory and implications for interaction and collaborative learning in computer conferences. *International Journal of Educational Telecommunications* 1 (1995), 147–166.
- [15] Ali Gürkan and Luca Iandoli. 2009. Common ground building in an argumentation-based online collaborative environment. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems* (France) (*MEDES '09*). Association for Computing Machinery, New York, NY, USA, Article 48, 5 pages. [doi:10.1145/1643823.1643882](https://doi.org/10.1145/1643823.1643882)
- [16] Enamul Hoque and Giuseppe Carenini. 2014. ConVis: A Visual Text Analytic System for Exploring Blog Conversations. *Computer Graphics Forum* 33 (06 2014). [doi:10.1111/cgf.12378](https://doi.org/10.1111/cgf.12378)
- [17] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) (*CHI '03*). Association for Computing Machinery, New York, NY, USA, 17–24. [doi:10.1145/642611.642616](https://doi.org/10.1145/642611.642616)
- [18] Luca Iandoli, Ivana Quinto, Anna De Liddo, and Simon Buckingham Shum. 2014. Socially augmented argumentation tools: Rationale, design and evaluation of a debate dashboard. *Int. J. Hum.-Comput. Stud.* 72, 3 (March 2014), 298–319. [doi:10.1016/j.ijhcs.2013.08.006](https://doi.org/10.1016/j.ijhcs.2013.08.006)
- [19] Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and Resolution on Wikipedia: A Case Study of Requests for Comments. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 74 (nov 2018), 24 pages. [doi:10.1145/3274343](https://doi.org/10.1145/3274343)
- [20] Dariusz Jemielniak. 2014. *Common Knowledge? An Ethnography of Wikipedia*. Stanford University Press, Standford, California.
- [21] Dariusz Jemielniak and Andreea Gorbatai. 2012. Power and status on Wikipedia. (2012).
- [22] Aniket Kittur and Robert E. Kraut. 2010. Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Savannah, Georgia, USA) (*CSCW '10*). Association for Computing Machinery, New York, NY, USA, 215–224. [doi:10.1145/1718918.1718959](https://doi.org/10.1145/1718918.1718959)
- [23] Christine Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017–2027. [doi:10.18653/v1/2021.eacl-main.173](https://doi.org/10.18653/v1/2021.eacl-main.173)
- [24] Christine Kock and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3824–3837. [doi:10.18653/v1/2022.emnlp-main.252](https://doi.org/10.18653/v1/2022.emnlp-main.252)
- [25] Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, consensus, coercion, control: cs*w or how policy mediates mass participation. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) (*GROUP '07*). Association for Computing Machinery, New York, NY, USA, 167–176. [doi:10.1145/1316624.1316648](https://doi.org/10.1145/1316624.1316648)
- [26] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) (*CSCW '12*). Association for Computing Machinery, New York, NY, USA, 265–274. [doi:10.1145/2145204.2145249](https://doi.org/10.1145/2145204.2145249)
- [27] Jonathan T. Morgan, Robert M. Mason, and Karine Nahon. 2011. Lifting the veil: the expression of values in online communities. In *Proceedings of the 2011 ICOnference* (Seattle, Washington, USA) (*iConference '11*). Association for Computing Machinery, New York, NY, USA, 8–15. [doi:10.1145/1940761.1940763](https://doi.org/10.1145/1940761.1940763)
- [28] Kevin K. Nam and Mark S. Ackerman. 2007. Arkose: reusing informal information from online discussions. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) (*GROUP '07*). Association for Computing Machinery, New York, NY, USA, 137–146. [doi:10.1145/1316624.1316644](https://doi.org/10.1145/1316624.1316644)
- [29] Chitu Okoli, Arvin Mesgari, Mohamad Mehdi, Finn Nielsen, and Arto Lanamäki. 2014. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology* 65 (12 2014), 2381–2403. [doi:10.1002/asi.23162](https://doi.org/10.1002/asi.23162)
- [30] Wenjing Pian. 2019. Factors affecting browsing duration on a health discussion forum: analysis of eye-tracking data. *Information Research* 24 (09 2019).
- [31] Xiaoxue Ren, Zhenchang Xing, Xin Xia, Guoqiang Li, and Jianling Sun. 2020. Discovering, explaining and summarizing controversial discussions in community QA sites. In *Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering* (San Diego, California) (*ASE '19*). IEEE Press, 151–162. [doi:10.1109/ASE.2019.00024](https://doi.org/10.1109/ASE.2019.00024)
- [32] Jodi Schneider, Alexandre Passant, and John Breslin. 2010. A Content Analysis: How Wikipedia Talk Pages Are Used. In *Proceedings of the Web Science Conference 2010 (WebSci '10)*. Raleigh, North Carolina, USA.
- [33] Jodi Schneider, Alexandre Passant, and John G. Breslin. 2011. Understanding and improving Wikipedia article discussion spaces. In *Proceedings of the 2011 ACM Symposium on Applied Computing* (TaiChung, Taiwan) (*SAC '11*). Association

- for Computing Machinery, New York, NY, USA, 808–813. doi:[10.1145/1982185.1982358](https://doi.org/10.1145/1982185.1982358)
- [34] Jodi Schneider, Alexandre Passant, and Stefan Decker. 2012. Deletion discussions in Wikipedia: decision factors and outcomes. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (Linz, Austria) (*WikiSym ’12*). Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. doi:[10.1145/2462932.2462955](https://doi.org/10.1145/2462932.2462955)
- [35] Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: how experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) (*CSCW ’13*). Association for Computing Machinery, New York, NY, USA, 1069–1080. doi:[10.1145/2441776.2441897](https://doi.org/10.1145/2441776.2441897)
- [36] John Short, Ederyn Williams, and Bruce Christie. 1976. *The Social Psychology of Telecommunications*. Pitman Press, London, UK.
- [37] Fernanda B. Viégas, Scott Golder, and Judith Donath. 2006. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) (*CHI ’06*). Association for Computing Machinery, New York, NY, USA, 979–988. doi:[10.1145/1124772.1124919](https://doi.org/10.1145/1124772.1124919)
- [38] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) (*CHI ’04*). Association for Computing Machinery, New York, NY, USA, 575–582. doi:[10.1145/985692.985765](https://doi.org/10.1145/985692.985765)
- [39] Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. 2007. Talk Before You Type: Coordination in Wikipedia. In *2007 40th Annual Hawaii International Conference on System Sciences* (*HICSS’07*). 78–78. doi:[10.1109/HICSS.2007.511](https://doi.org/10.1109/HICSS.2007.511)
- [40] Joseph Walther and Jeong-woo Jang. 2012. Communication Processes in Participatory Websites. *Journal of Computer-Mediated Communication* 18 (10 2012). doi:[10.1111/j.1083-6101.2012.01592.x](https://doi.org/10.1111/j.1083-6101.2012.01592.x)
- [41] Joseph Walther and M. Parks. 2002. Cues Filtered Out, Cues Filtered In: Computer-Mediated Communication and Relationships. (01 2002).
- [42] Wikipedia. 2024. Administrators' noticeboard. Retrieved July 25, 2024 from https://en.wikipedia.org/wiki/Wikipedia:Administrators%27_noticeboard
- [43] Wikipedia. 2024. Dispute resolution requests/Noticeboards. Retrieved April 11, 2025 from https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution_requests/Noticeboards
- [44] Wikipedia. 2024. Five pillars. Retrieved Oct 18, 2024 from https://en.wikipedia.org/wiki/Wikipedia:Five_pillars
- [45] Wikipedia. 2024. Harassment. Retrieved Mar 25, 2025 from <https://en.wikipedia.org/wiki/Wikipedia:Harassment>
- [46] Wikipedia. 2024. No personal attacks. Retrieved Mar 25, 2025 from https://en.wikipedia.org/wiki/Wikipedia>No_personal_attacks
- [47] Wikipedia. 2024. Requests for adminship. Retrieved July 25, 2024 from https://en.wikipedia.org/wiki/Wikipedia:Requests_for_adminship
- [48] Wikipedia. 2024. The three-revert rule. Retrieved July 25, 2024 from https://en.wikipedia.org/wiki/Wikipedia>Edit_warring#The_three-revert_rule
- [49] Wikipedia. 2024. Wikipedia:BOLD, revert, discuss cycle. Retrieved July 25, 2024 from https://en.wikipedia.org/wiki/Wikipedia:BOLD,_revert,_discuss_cycle
- [50] Wikipedia. 2024. Wikipedia:Edit warring. Retrieved July 25, 2024 from https://en.wikipedia.org/wiki/Wikipedia>Edit_warring
- [51] Wikipedia. 2025. Dispute resolution requests. Retrieved April 11, 2025 from https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution_requests
- [52] Wikipedia. 2025. Edit count. Retrieved Mar 25, 2025 from https://en.wikipedia.org/wiki/Edit_count
- [53] Taha Yasseri, Robert Sumi, András Rung, Andras Kornai, and János Kertész. 2012. Dynamics of Conflicts in Wikipedia. *PLoS one* 7 (06 2012), e38869. doi:[10.1371/journal.pone.0038869](https://doi.org/10.1371/journal.pone.0038869)
- [54] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 196 (Nov. 2018), 27 pages. doi:[10.1145/3274465](https://doi.org/10.1145/3274465)
- [55] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (*CSCW ’17*). Association for Computing Machinery, New York, NY, USA, 2082–2096. doi:[10.1145/2998181.2998235](https://doi.org/10.1145/2998181.2998235)

A Creating the Technology Probe

In the following, we describe how we worked to select possible disputes from talk pages and how we iterated through LLM prompts to create a discussion summary as a type of technology probe to

be used in our research. Once we had a prompt that could create a reasonably reliable discussion summary, that LLM prompt was applied to a selected set of talk page disputes.

A.1 Selecting Talk Page Disputes

An ideal discussion for our study needed to involve debate about the article content and have the potential to escalate into a dispute rather than proceeding smoothly. We identified controversial discussions by noting that Wikipedians cite policies during their debates [25]. We first compiled a list of policies that potentially imply conflict. These included foundational guidelines for article content such as "Neutral point of view," "No original research," and "Verifiability," as well as those implying contentious debate such as "Edit warring" and the "Three-revert rule."⁴ Next, we used the "What links here" function, which provides a list of every page that cites these policies. By applying the namespace filter to isolate article talk pages, we obtained a targeted list of talk pages with discussions referencing these policies.

From the discussions identified through this process, we selected discussions where there appeared to be a disagreement (a dispute) and which met some additional complexity criteria. We chose discussions involving more than three editors, posting differing viewpoints, and with more than 10 comment posts. We excluded arguments revolving around simple yes or no questions, such as whether to delete a sentence. We also ignored any discussions with polling or "not vote" decision making. When selecting the discussions we considered how our LLM prompt might apply. We focused on selecting disputes that focused on textual content within the article. Consequently, discussions about visual and structural components like images, infoboxes, more or fewer headings, or similar features were excluded.

This selection process resulted in a set of 24 candidate discussions from 22 articles. We used the 24 candidates to refine our prompting and summary generation. The subjects of the articles were: 2003 invasion of Iraq, Abortion, Black Lives Matter, Black supremacy, Domestic violence, Anti-Americanism, Adolf Hitler, Masculism, NATO, Terrorism, Communism, Capitalism, Compiler, Erotica, February 29, Race, French Language, Barber paradox, BlackBerry, Democratic Republic of Afghanistan, Floppy disk, and Dallas.

A.2 Iterative Prompt Refinement

We used ChatGPT 3.5 to generate the summaries.

Our early prompts focused on what we might be able to get an LLM to do with generic text from a Wikipedia talk page. In this initial stage, we provided the discussion text alone. The prompt labeled the text as a "discussion" and asked the LLM: "What is the discussion talking about;" and "What is the discussion suggesting;" Responses generated in this iteration failed to accurately identify the main points of the discussion.

Consequently, we expanded the prompt to specify that the text was a discussion from an online community and directed the LLM to focus on identifying the primary discussion topic. An example prompt for the second iteration included: "This is a discussion in an online community. [discussion] What are they discussing? What is the main topic of the discussion;" In this prompt "[discussion]" was replaced with the text of the selected talk page discussion. The LLM responses illustrated a lack of understanding of the article content and discussion context, but improved some with regard to the primary focus of the discussion itself.

Therefore, the prompt for the third iteration was refined to: "This is the content in question. [content] This is a discussion in an online community on the content. [discussion] What is the main

⁴"Three-revert rule" stipulates that an editor must not perform more than three reverts on a single page within a 24-hour period. This serves as the benchmark for blocking or banning users engaged in edit wars [48].

topic of the discussion? Only explain the main topic of the discussion.“ In this prompt “[content]” was replaced with the text of the specific section that the editors were discussing, and, like before “[discussion]” was replaced with the talk page discussion text. This version better captured the discussion topic, particularly in cases where the editors addressed specific edit details directly without mentioning the article subject and the related subsection in the article.

For the fourth iteration, we developed three slightly different prompts. We revised the prompt to specify that the discussion was about Wikipedia article content. The three prompts were devised, each with subtle differences in emphasis on the discussion being a dispute between editors and on conciseness of the result:

- (1) “This is the content in question. [content] This is a discussion containing a dispute about what to write in a Wikipedia article. [discussion] Identify the dispute. Be very concise.”
- (2) “This is the content in question. [content] This is a discussion about what to write in a Wikipedia article. [discussion] Explain the main topic of the discussion. Be very concise.”
- (3) “This is the content in question. [content] This is a discussion about what to write in a Wikipedia article. [discussion] Concisely explain the main topic of the discussion. Be very concise.”

The first prompt explicitly instructed the LLM to identify the dispute within the discussion, the second prompted for identification of the main discussion topic, and the third emphasized conciseness, twice.

In our fifth iteration, we evaluated the three prompts using the 24 discussions we had collected. The first prompt typically yielded straightforward and clear answers without excessive length, although occasionally conveyed a stronger meaning than the original discussion. The second prompt generally produced the most detailed responses, but these were sometimes overly lengthy or abstract. Conversely, the third prompt usually generated the most concise answers but occasionally resulted in incorrect summaries of the arguments. Consequently, we decided to employ the first prompt for generating our probe.

A.3 Technology Probe Summaries

For the final technology probe summaries, we selected four discussions from the articles Barber Paradox, February 29, French Language, and Erotica. These subjects were chosen to represent a range of domains while avoiding highly contentious or polarizing subjects that might trigger strong personal biases or emotions and interfere with understanding general cognitive processes involved in reading. We also ensured that each discussion was sufficiently complex yet concise enough for participants to review and complete the tasks within the interview time.

Barber Paradox. The dispute revolves around edits to the Wikipedia article on the Barber paradox. The disagreement centers on whether stating that the barber does not exist is a valid resolution to the paradox. One editor argues that claiming the nonexistence of the barber violates a premise of the paradox and is not a solution. The other editor suggests that if the lack of existence is not a valid resolution, the section should be reworded to convey that there is no solution rather than reverting the changes. The dispute also involves concerns about the use of established mathematical facts versus perceived fringe theories in the article.

February 29. The dispute revolves around the inclusion of information about legal age determination for individuals born on February 29 in the Wikipedia article. [Orange] wants to include specific details about legal age calculations on leap years, arguing that it is relevant to the uniqueness of February 29. [Lemon] suggests placing the information in the Legal age article, while others recommend discussing and seeking consensus on the Wikipedia talk:WikiProject Days of the

year/Template. The disagreement highlights the challenge of finding an appropriate place for specific content within Wikipedia's structure and the need for consensus among editors.

French Language. The dispute revolves around whether French should be considered an official language of the Vatican City State. [Kumquat] argues that French is one of the official languages, citing sources such as French Wikipedia, Vatican News, and the French government's diplomatic website. However, [Banana] and [Blueberry] dispute this, stating that the sources provided are not from official Vatican sources and that other sources indicate that only Italian is the official language of the Vatican City State. [Blueberry] also raises concerns about the conflation of official language and working language in the article and argues that French is used in diplomacy but not officially recognized. The disagreement involves interpretations of sources, the distinction between official and working languages, and whether consensus has been reached on the matter.

Erotica. The dispute revolves around whether a section on "Student erotica" should be included in the Wikipedia article. Some users argue that it is too narrow and recent a genre to be added to a broad article like "Erotica," suggesting that it deserves its own article. Others contend that the section is relevant, sourced, and should stay in the article until it grows enough for its own article. The discussion involves multiple edits and disagreements on the appropriateness of including "Student erotica" in the general topic article.

A.4 Participant-Generated Summaries

Received October 2024; revised April 2025; accepted August 2025

| ID | Topic | Summary |
|-----|-----------------|---|
| P01 | Barber Paradox | IP editor brings a complaint about their edits being reverted without reason. IP's edits concerned adding content regarding a potential "solution" to the paradox which is the subject of the article. Another editor joins the discussion and explains that IP's edits reduced the quality of the article in question and cited 3RR, indicating that IP had been reverting edits consecutively and without subsequent discussion. A compromise was reached and the solution was implemented by the registered editor involving a more explicit explanation of the initially perceived problem. |
| P06 | Barber Paradox | IP address attempted to make an improvement, apparently more than once, and was reverted each time by a registered editor who felt that the change wasn't improvement, and explained why. IP address didn't agree with that rationale, and suggested rewording rather than reverting. Registered editor made an edit that might satisfy both sides, but IP address still thinks something is missing. Both editors are assuming good faith and civil. |
| P10 | Barber Paradox | An argument is being made that content was removed because it was "not an improvement". An editor is raising concern that the revert reason is not appropriate. |
| P11 | Barber Paradox | IP editor added good faith contributions citing a educational institution. They should have attributed their statements to that institution/professor, but they didn't need to be reverted. A registered editor defended the revert saying that the contributions were not helpful. IP editor asks if the registered editor would re-write a section that has been removed in a better way, unanswered. |
| P13 | Barber Paradox | This discussion is a good discussion and very educative. Educative in a sense that, all parties stated Wikipedia policies to back their 'argument' and later the discussion reached a consensus where both parties were happy. |
| P14 | Barber Paradox | One editor added an unsourced addition that another editor then removed, and a different editor defended the removal. I also concur with the removal. |
| P04 | February 29 | Two editors violated the three-revert rule (WP:3RR), so an uninvolved editor brought it up on the talk page. A solution involving putting the disputed material in the article body rather than the body was proposed, but one of the editors involved in the edit war rejected the proposed solution in favor of their own solution. By the end of the discussion, no clear consensus was established for the editor's change. |
| P05 | February 29 | [Grape] is trying to resolve an edit war between [Orange] and [Lemon] about whether the article should have a section on the date's interaction with one's legal age. [Orange] is for its inclusion; [Lemon] is against the inclusion. |
| P07 | February 29 | This discussion follows several edits and reverts on an article page. A cool-headed editor reminds the "breaking"-editor who has created the editing issue that instead of continuing with the edit-revert-edit-revert pattern, which can lead to getting blocked, the better way to deal with differences regarding article content is to engage in discussion on the talkpage as the goal is to reach consensus. The breaking-editor is frustrated, and suggests "breaking" a template. This suggestion is clearly not held by others as it doesn't lead to consensus. A third editor suggests shifting the content in question to another page, and a specific page is linked. A fourth editor tries to conclude the conversation in terse, short sentences, basically reminding the editor regarding the goal of consensus. |
| P02 | French Language | One user, [Apple], thought that France was an official language of Vatican city. However, he cited a Wikipedia article, which (although in another language) is still not counted as a credible source. The [Apple] user had two other non-Wikipedia sources, but he still lost a bit of credibility just by citing Wikipedia. Even though he still backed up his claims, he admits to being involved in edit wars and even goes as far as to revert another user's edit, indicating a bit of defensiveness. *At the end of the discussion, the [Apple] user was revealed to be a sockpuppet, so he could've been a troll. |
| P03 | French Language | In this discussion, editors discussed whether or not this article should state that French is an official language of the Vatican City. An editor in support of this ([Apple]) posted various sources that they said supported the inclusion of the Vatican as a Francophone country, including Vatican News, the French Wikipedia's article on the language of the Vatican, and Diplomatique. The user said that international institutions (such as the Francophonie) recognize the Vatican as a Francophone country. Other users, however, were unconvinced. One ([Blueberry]) pointed out that La Francophonie does not appear to consider the Vatican to be a Francophone country (it is not a member nor observer of that institution), and that the sources presented by [Apple] were unreliable (as user-generated) or did not actually support the claims that the Vatican uses French as an official language (the user pointed to specific language in the sources to support this). Another user ([Banana]) said that basically all sources refer to the language as Italian, though didn't specifically list sources. Following this discussion, [Apple] was blocked as a sockpuppet, and their comments struck as such. No consensus was obtained to add the Vatican City as being a country with a French official language. |
| P08 | French Language | The crux of the dispute is that French was added as an official language, citing the French diplomatic website as well as a source from Vatican News listing it as "an official diplomatic language". Some of the sources used were questioned and an editor opposing the addition of French stated that it was an official diplomatic language but not an official language. The discussion ended after the editor attempting to add the Vatican City was blocked as a sockpuppet. |
| P09 | French Language | User [Apple] suggested listing Vatican as one of the countries in which French is an official language and gave three sources: French Wikipedia, a website by the government of France, and a Vatican news website. Users [Banana] and [Blueberry] rejected the French Wikipedia as a circular source, and, while they saw the other two sources as more valid, they disagreed that they prove that French is an official language of the Vatican and said that it is, at most, a diplomatic or a working language, which is not the same as an official language. User [Carrot] struck out [Apple]'s text saying that that is a sockpuppet account. |
| P12 | French Language | The discussion here is about whether or not French is an official language of the Vatican. It has two elements, what constitutes an official language and what are reliable sources. |